· Since probabilities depend on weights (minimizing weighted MSE) and weights depend on probabilities we use
- iterative weighted least squares / Fisher scoring to decide weights/probs (ends up being the same as gradient descent)

Multidimensional Probability:

Have random vector $\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$, $Pr(\vec{x} \in A) = \int_A f(\vec{x}) d\vec{x} = \int_A f(x_1, \dots, x_p) dx_1 \dots$

· probability that $\vec{X}$ is within $\pm h$ of $\vec{x} \approx h^p f(\vec{x})$

$E[\vec{X}] = \int \vec{x} f(\vec{x}) d\vec{x} = \begin{bmatrix} E[X_1] \\ \vdots \\ E[X_p] \end{bmatrix}$, $Var[\vec{X}] = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & Cov(X_1,X_p) \\ Cov(X_2,X_1) & \ddots & \vdots \\ Cov(X_p,X_1) & \cdots & Var(X_p) \end{bmatrix}_{p \times p}$

$Var[\vec{X}]$ is positive semidefinite: $\vec{u}^T Var(\vec{X}) \vec{u} \geq 0 \quad \forall \vec{u}$
and it is symmetric.
These indicate it has an eigendecomposition:
$Var[\vec{X}] = w d w^T$ where d = diagonal matrix of eigenvalues of $Var[\vec{X}]$ (all eigenvals are ≥ 0)
w = $p \times p$ matrix whos columns are eigenvectors of $Var[\vec{X}]$, scaled so that each column has length 1. All eigenvectors are orthogonal to each other so $w^T = w^{-1}$

For vector $\vec{a}$, $Var[\vec{a}\vec{X} + \vec{b}] = \vec{a}\, Var[\vec{X}]\, \vec{a}^T$

$$E[d^{-1/2} w^T(\vec{x} - E[\vec{x}])] = \vec{0}$$

$$Var[d^{-1/2} w^T(\vec{x} - E[\vec{x}])] = I$$

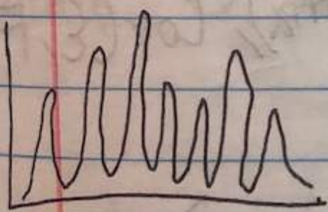
- cond independence
- nonparametric model needs to have more args when more data points are added

Emperical probabilities:

$$\hat{P}(X \in A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \in A\}$$

Kernel Density Estimation: (estimating probability densities for data features)

$$\hat{f}_{kern}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$



- large h means less peaks that are wider
- lower h means more and higher peaks

- Use cross validation to pick bandwidth with either $\int (f(x) - \hat{f}(x))^2 dx$ (integrated squared error) or $\int -\log(\hat{f}(x)) f(x) dx$ (log likelihood)

(with IID samples for large n, by central limit theorem

$$\int -\log(\hat{f}(x)) f(x) dx \approx \frac{1}{n} \sum_{i=1}^{n} -\log \hat{f}(x_i)$$

- If we can find some X_j such that $X_1, \ldots, X_p$ are all conditionally independent given X_j, then $f(X_1, \ldots, X_p) = f(X_j) \prod_{i=1} f(X_i | X_j)$. This reduces number of params we need to estimate (and help lessen curse of dimensionality).

Factor Model:

- Assume $E[\vec{X}]=0$
- Assume there is a q-dimensional random vector $\vec{F}$ with $E[\vec{F}]=0$, $Var[\vec{F}]=I$, $q<p$
- $\vec{F}$ is latent/unobserved/hidden, $\vec{X}$ is manifest/observable
- Assume $\underset{p\times1}{\vec{X}} = \underset{p\times q}{w}\,\underset{q\times1}{\vec{F}} + \underset{p\times1}{\vec{\varepsilon}}$, with $E[\vec{\varepsilon}]=0$, $Var(\vec{\varepsilon})=\Psi$ (Ψ = diag matrix), $Cov(\vec{\varepsilon},\vec{F})=0$

So under these assumptions

$$Cov(X_i, X_j) = Cov\left(\sum_{k=1}^{q} w_{ik}F_k + \varepsilon_i, \sum_{l=1}^{q} w_{jl}F_l + \varepsilon_j\right)$$

$$= Cov\left(\sum_{k=1}^{q} w_{ik}F_k, \sum_{l=1}^{q} w_{jl}F_l\right)$$

$$= \sum_{k=1}^{q}\sum_{l=1}^{q} w_{ik}w_{jl}\, Cov(F_k, F_l)$$

$$= \sum_{k=1}^{q} w_{ik}w_{jk} \quad \text{(since } Cov(F_k,F_l) \text{ is 0 when } k\neq l \text{ and 1 when } k=l\text{)}$$

So X_i and X_j are correlated when they load on the same factors